Estimating the Survival Distribution of Aluminum Processing Pots

by

Emily L. Butler

Honors Project

Project Advisor: Joel Greenhouse, Department of Statistics

Presented to the Department of Statistics and the Dean's Office
of the College of Humanities and Social Sciences in
Partial Fulfillment of the Requirements for the H&SS Senior Honors Program

Carnegie Mellon University

College of Humanities and Social Sciences

May 2011

Abstract:

The goal of this thesis is to specify a probability model for time-to-failure items in a manufacturing process. Specifically, I am interested in the time-to-failure of containers, called pots, in which aluminum is produced. Aluminum smelting is a very complex and sensitive process. The process uses specialized large carbon lined steel pots that contain a carbon rod and a molten cryolite bath, in which the final product aluminum is produced. A problem arises when the pots fail, for example, when a pot is unable to operate at a certain temperature the molten aluminum hardens resulting not only in wasted product, but also wasted time and resources to clean and remove the pot. In this thesis, I investigate different parametric models for the time-to-failure distribution for aluminum pots: the Weibull model, the Weibull change-point model, the Gompertz model, and the Gompertz-Makeham model. My work involves understanding the structure of the data, specifying which distributions to investigate and why, estimating parameters using maximum likelihood estimation, visualization methods for time-to-failure data, and how to test the fit of the models. These topics will all be discussed using a data set provided by Alcoa, Inc.

Chapter 1: Introduction

The goal of this thesis is to specify a probability distribution for the time-to-failure of items in a manufacturing process. Specifically, I am interested in the time –to-failure of the containers, called pots, in which aluminum is produced. Discussion will involve understanding the structure of the data, specifying which distributions to investigate and why, estimating parameters using maximum likelihood estimation, how to fit the model, visualization methods for time-to-failure data, and how to test the fit of the models. These topics will all be discussed using a data set provided by Alcoa, Inc.

Aluminum Production

Aluminum is frequently used in automobile manufacturing, aerospace engineering, construction, and for many household uses. It has become an essential part of everyday life. However, aluminum is not found naturally in the environment; it needs to be altered from its natural state. The element "aluminum" is too reactive to occur by itself in nature—instead, it combines with other elements to form various minerals ("How Aluminum is Produced"). The most abundant source of aluminum is located in "bauxite ore" and, consequently, it is a crucial component in aluminum manufacturing. There are two major steps in the production of aluminum. First, alumina is extracted from bauxite ore though the Bayer Process, then it undergoes smelting to convert the alumina to aluminum, using the Hall-Héroult Process.

During the Bayer Process, the bauxite ore is crushed and mixed with a mild sodium hydroixide solution. The mixture is placed in a digester where it faces high temperatures and extreme pressure, resulting in dissolved aluminum oxide and other residue (including silicon, lead, titanium, etc, which sink to the bottom of the digester). After the water is evaporated out of

the aluminum oxide mixture and the solution is condensed, it crystallizes, forming aluminum hydroxide and sodium hydrizide, also known as alumina.

Following the Bayer process is the Hall-Héroult Process. The smelting process occurs in a number of large carbon lined steel reduction pots which contain a carbon rod and a molten cryolite bath (sodium aluminium fluoride). While the alumina is mixed with a cryolite bath, an electrical current runs through that mixture from the positively charged carbon rod to the negatively charged carbon lined pots (at about 5.25 volts). As the electrical current flows through the mixture, carbon is combined with oxygen in the alumina, producing aluminum and carbon dioxide as a byproduct. The molten aluminum settles to the bottom of the pot while the carbon dioxide is released through the top. The molten aluminum is then siphoned off where it can be collected and turned into the various alloys used in everyday products ("Aluminum Smelting and Refining"). The entire smelting process requires rows of reduction pots, or potlines, be in production 24 hours a day, 365 days a year. It is difficult to stop and start the smelting process because the result is a loss of money, energy, and product. Furthermore, if the temperature of the pots decreases and the molten aluminum hardens, the repair and clean up is costly and time consuming. Unfortunately, as difficult as it is to change or repair the reduction pots, these pots do not last forever.

It is very difficult to estimate when a pot stops working. Being able to estimate when a pot stops performing efficiently would not only save a company time, money, and energy but also reduces the costs to consumers. This problem was brought by Alcoa Inc.

Alcoa provided data from 131 pots. Of these pots, 47 had "failed," or were unable to continue producing aluminum efficiently, and 84 were still operating at the time the data finished being collected, which was December 31, 2009. The goal of this research is to find a probability

model for the pot failure times. I will use statistical models and tests based on methods for the analysis of survival data to best describe the distribution of the failure time of these aluminum smelting pots.

Survival analysis uses information about the failure time of the pots to estimate the time-to-failure distribution of the pots. The purpose to this analysis is to discover which survival distribution these data follow. These results will be useful to eventually create a predictive model for pot failure. At the onset it will be useful to review what survival analysis is, why it is germane for this type of problem, and what kind of information it provides.

Survival Analysis

In many investigative fields, including engineering and medical research, researchers are interested in estimating the time until an event of interest occurs. In a field involving live subjects like the medical field, one may be interested in estimating the time until death of the subject from the beginning of observation time, for example birth, onset of a disease, entrance or start of a clinical trial. In an engineering problem such as the one I am concerned with, the focus is on time until failure of an object, such as a pot or a lightbulb. My analysis is interested in estimating the survival time, or how long the pots function until they fail. The collection of methods used for the analysis of time-to-failure data is called survival analysis.

Survival analysis is the study of lifetime distributions, in this case, the lifetime of the pots where lifetime refers to the period from the beginning of observation until failure. The time when the pots fail is referred to as the failure time. In our case, the failure time will be day the pot stopped working.

Before any conclusions are able to be drawn or before it's possible to make predictions, it is necessary to specify a probability model for the distribution of the survival times. I will be

using familiar representations of probability distributions, the probability density function and the cumulative distribution function, as well as three unfamiliar functions that are typically used in survival analysis: the survival function, the hazard function and the cumulative hazard function. I will show how each of these functions are related such that if one function is known, it is possible to derive the others. Each of the functions highlights different features of the distribution of survival times.

To show these relationships, start by assuming that random variable T is continuous, T>0, and has a probably density function (pdf), $f(t)$. The cumulative distribution function (cdf), $F(t)$, can be obtained by integration the pdf:

$$F(t) = \int_0^t f(y)dy$$

which can also be denoted as $P(T < t)$, the cumulative probability of the occurrence of the random variable T up to a given point, t. In words, the cumulative distribution function measures the cumulative probability that an object fails before time t ("Related Distributions"). In survival analysis, the interest is in the probably of survival beyond time t, or

$$S(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(y)dy$$

which is called the survival function. It is possible to derive the pdf from the survival function by taking the negative derivative of the survival function with respect to t, or

$$f(t) = \frac{-dS(t)}{dt}$$

To describe the lifetime distribution of a random variable, it's also possible to use the hazard function, $h(t)$, which measures the instantaneous failure rate at time t. The hazard function is

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\Delta \to 0} \frac{P(t < T \leq t + \Delta)}{\Delta P(T > t)}$$

The hazard function is the risk of failure in a small time interval, given survival at the beginning of the time interval (Elandt-Johnson and Johnson, 60-63). As a function of time, a hazard function may be increasing, meaning as time increases the rate for failure increases, for example, when a patient is untreated for a disease such as cancer; may be decreasing, for example, as a person is recovering from severe trauma like a surgery; or may be constant, meaning the rate of failure is the same regardless of how much time has passed.

The cumulative hazard function, or the accumulation of hazard over time, can be found by integrating the hazard function from 0 to t,

$$H(t) = \int_0^t h(y)dy$$

or equivalently can be found by taking the negative logarithm of the survival function (Smith, 3-13):

$$H(t) = -\log S(t)$$

The hazard function will be discussed in more detail in the next chapter. As one can see from the table below, it is possible to find any of these functions if one of the other functions is known.

Probability Density Function: $f(t)$	$f(t) = F'(t) = 1 - S'(t) = S(t)h(t)$
Cumulative Density Function: $F(t)$	$F(t) = \int_0^t f(y)dy$
Survival Function: $S(t)$	$S(t) = 1 - F(t) = e^{-H(t)}$
Hazard Function: $h(t)$	$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1-F(t)}$.
Cumulative Hazard Function: $H(t)$	$H(t) = -\log S(t) = \int_0^t h(y)dy$

As an example, consider a random variable T with an exponential probability distribution with parameter θ

$$f(t) = \frac{1}{\theta} e^{-\frac{1}{\theta}t}.$$

By integrating, the cdf of T is

$$F(t) = \int_t^0 \frac{1}{\theta} e^{-\frac{1}{\theta}y} dy = 1 - e^{-\frac{1}{\theta}t}.$$

The survival function is then

$$S(t) = 1 - (1 - e^{-\frac{1}{\theta}t}) = e^{-\frac{1}{\theta}t}$$

and the hazard function is

$$h(t) = \frac{\frac{1}{\theta} e^{-\frac{1}{\theta}t}}{e^{-\frac{1}{\theta}t}} = \frac{1}{\theta}$$

From the hazard function, the cumulative hazard function is

$$H(t) = \int_0^t \frac{1}{\theta} dt = \frac{1}{\theta} t$$

The survival function is central in survival analysis and has three important properties:

1. It is a monotonically decreasing function, which is logical because it is the complement of the cdf, which is monotonically increasing;
2. $S(0) = 1$
3. $S(\infty) = 0.$

In words, at time $t = 0$ all pots are working, meaning the survival function equals 1 (no pots have failed yet), and as $t \to \infty$, eventually all of the pots will fail which means the survival function will eventually equal 0 (Lee and Wang, 10-12).

A complication of survival data is that it is often not possible to observe all of the pots until they have failed. However, it is important to have information about which pots failed and

which ones did not. The pots that have not failed by the end of the observation period are referred to as censored observations. Therefore, the pot data consists of time until either failure or time until censored (Smith, 73-77). We will use an indicator variable to indicate if the pot was a failure or if it was censored.

The statistical challenge is given a data set to find the best fitting probability model for the distribution of survival times. For a parametric survival model, this means estimating the parameters and then assessing the goodness-of-fit of the model. Because of the one-to-one relationship among the different representations within a family of survival distributions, once the parameters of the distribution are estimated it's possible to display $f(t), S(t), h(t)$, or $H(t)$. Maximum likelihood estimation will be used to estimate the parameters and the likelihood ratio test and AIC will be used to assess the fit of different survival models. I will also use Kaplan-Meier nonparametric estimates of $S(t)$ and of $H(t)$ for exploratory visualization of the survival data and to help assess the fit of the models.

Graphing the survival function is important because it provides valuable insight into the behavior of the data. The Kaplan-Meier estimate of the survival curve will remain flat until there is a failure at which time the curve will drop an amount proportional to how many items failed at that time. When there are no more failures, the curve will flatten out again until it reaches another failure, where again the drop of the curve with be proportionate to the amount of pots that have failed. This means that the steeper the curve, the more failures there are and the larger the hazard rate. If the curve is relatively flat and has a shallow slope, there are a lot of pots surviving, i.e. pots are failing at a slow rate. The Kaplan-Meier estimate of the survival curve does not depend on any parametric assumptions about the underlying probability distribution of the data (Smith, 96-98).

An Example

To illustrate the different functions used in the analysis of survival data, I have generated data from an exponential distribution where the parameter θ equal one. Imagine that the data were collected on how long a light bulb lasts until it burns out. To do this, 30 light bulbs were turned on at the same time and each one is observed until it burns out, or fails. The survival times are found in the table below.

Survival Time of 30 Lightbulbs in Months														
0.020	0.025	0.059	0.062	0.145	0.186	0.196	0.197	0.205	0.210	0.262	0.314	0.511	0.604	0.678
0.695	0.740	0.760	0.846	0.86	0.914	0.992	1.181	1.194	1.309	1.995	2.255	2.509	2.910	5.543

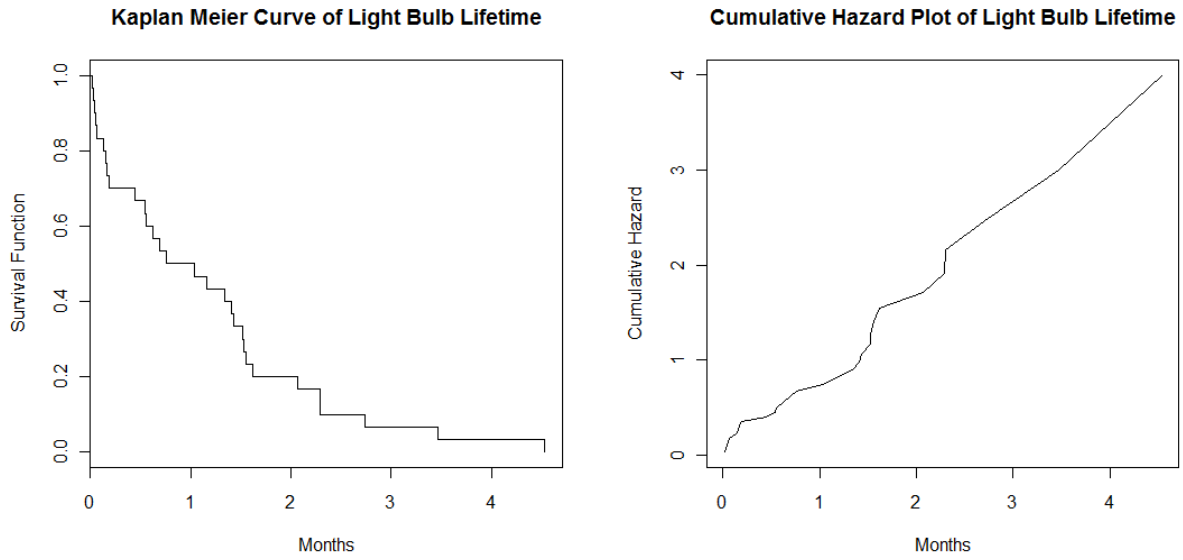
To estimate the survival function, $S(t)$, simply count up the observations larger than time t The easiest way to do this is create a latent variable, Z, where

$$Z = \begin{cases} 1 \; if \; T > t \\ 0 \; if \; T \leq t \end{cases}$$

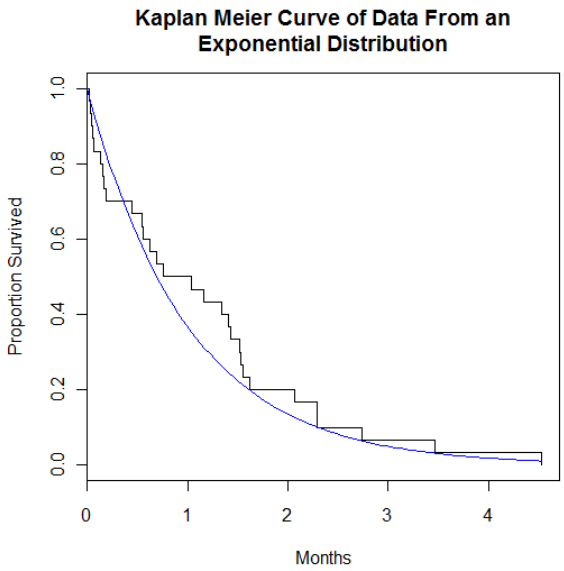
If there are no censored observations the empirical survival function is

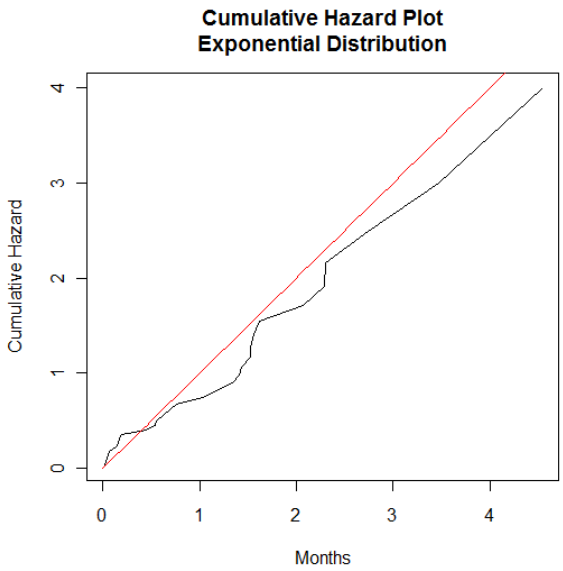
$$S_n(t_i) = \frac{\#\; observations > t}{n} = \frac{1}{n}\sum_{i=1}^{n} Z(t_i)$$

where t_i represents the observed survival times. To calculate $S(0)$, all 30 light bulbs were burning at time equals 0, therefore $S(0) = \frac{30}{30} = 1$. Similarly, at time ∞, zero light bulbs are still working so $S(\infty) = 0$. Therefore, these two conditions are happily satisfied. A plot of the estimate of the empirical survival function, that is, the Kaplan-Meier curve is on the next page on the left. The plot of the empirical cumulative hazard function is on the right. The empirical cumulative hazard function is found by taking – log(Kaplan-Meier Curve) at each time point.



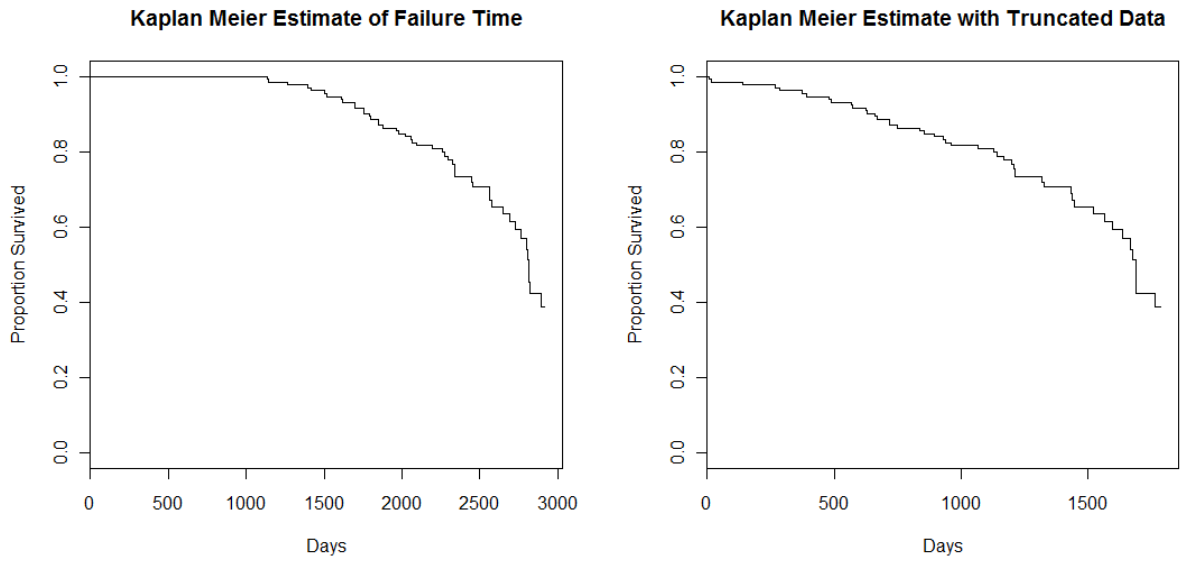
To investigate the underlying distribution of the data, one can fit various survival functions to the data and visually compare how similar the survival functions are to the Kaplan-Meier estimate of the survival function. For example, it's hypothesized that this data is exponentially distributed. To test this hypothesis, plot the survival function of the exponential distribution on the Kaplan-Meier plot, and plot the hazard function of the exponential distribution with parameter $\theta = 1$ on the cumulative hazard plot. As one can see from the plots on the next page, exponential survival curve appears to be similar to the Kaplan-Meier empirical survival curve. However, even though the data were generated from an exponential distribution with $\theta = 1$, because of sampling variability we see deviations of the observed data from the theoretical model, especially in the upper tail.


Kaplan Meier Curve of Data From an
Exponential Distribution
Proportion Survived
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
Months

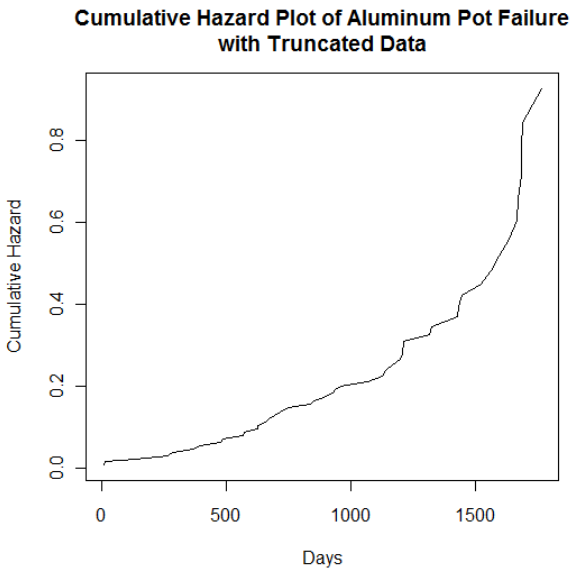

Cumulative Hazard Plot
Exponential Distribution
Cumulative Hazard
0
1
2
3
4
0
1
2
3
4
Months

Chapter 2: Fitting Distributions

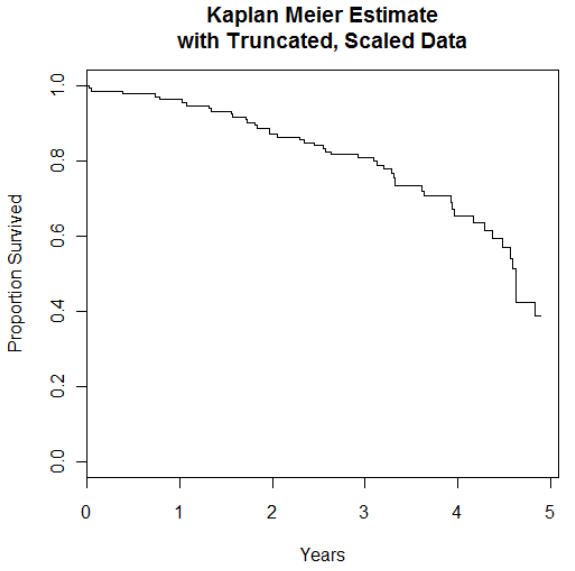
Recall that the goal of this analysis is to specify a probability model for the failure times of aluminum smelting pots using survival analysis. Before investigating which distribution fits best, it is important to become familiar with the structure of the data. Below are two Kaplan-Meier curves of the underlying survival function. On the left-most plot, note an obvious flat horizontal line from 0 to 1139 days, indicating that no pots failed within the first 1139 days. Because there is such an extend time until the first failure occurs, I will truncate the data prior to day 1138 to make the analysis conditional on an initial period in which there are no failures. The conditional Kaplan-Meier survival curve is shown below on the right most plot. Subsequently, when I refer to the survival distribution I mean the conditional survival distribution, $S(t|T > 1138)$.

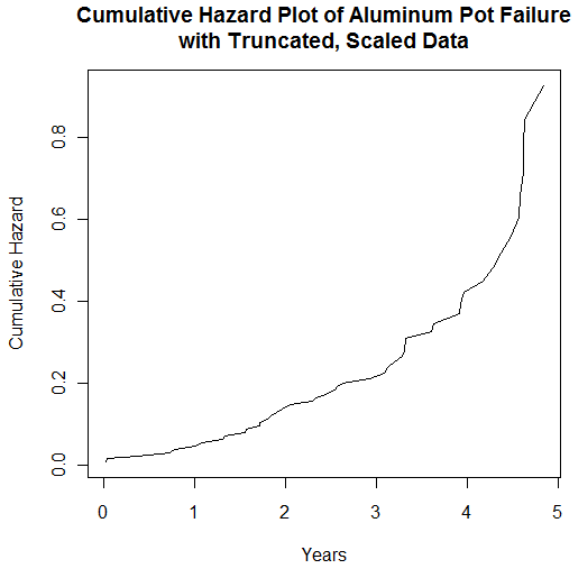


The plot on the following page is the cumulative hazard plot for the conditional survival data.



Even after the data is truncated, the range of failure times is still quite large, ranging from 0 to 1500 days. To facilitate numerical calculations and for ease of interpretation, I have transformed the data to years. As is visible in the plots of the survival and cumulative hazard functions below, the only thing that has changed is the time scale; the shape and behavior of the curves are the same. For the rest of this thesis, time will be reported in years.





Maximum Likelihood EstImation

Before any distribution can be fit to the data, the parameter values need to be estimated. For example, the survival function for the Weibull distribution is

$$S(t) = \exp(-(\theta t)^{\gamma})$$

where θ and γ are unknown parameters and t is the failure time. Different families of probability distributions have different unknown parameters. Given a family of probability distributions, the member of that family that best fits a given data set is found by estimating the parameters. The method of choice for parameter estimation is maximum likelihood estimation (MLE).

For survival data, the likelihood function is

$$L(\theta) = \prod_{i=1}^{n}[f(t_i,\theta)]^{\delta}[S(t_i,\theta)]^{1-\delta},$$

where θ is the parameter of interest, $f(t_i)$ is the pdf of probability distribution, $S(t_i)$ is the survival function, t_i is the failure times of the ith pot, for i=1…131, and δ is the status of the pot, i.e. whether it is censored ($\delta = 1$) or not ($\delta = 0$). The contributions to the likelihood function for the pots that have failed ($\delta = 1$) is $f(t_i)$ and for the pots that have not failed ($\delta = 0$) is $S(t_i)$ ("Maximum likelihood estimation").

Working with the logarithm of the likelihood function is often easier. Because a logarithm is a monotone transformation, the values of the parameters that maximize the likelihood function also maximize the log likelihood function. Because$f(t_i) = h(t_i)S(t_i)$, the likelihood function can be written:

$$(1)\quad L(\theta) = \prod_{i=1}^{n}[f(t_i,\theta)]^{\delta_i}[S(t_i,\theta)]^{1-\delta_i} = \prod_{i=1}^{n}[h(t_i,\theta)S(t_i,\theta)]^{\delta_i}\,[S(t_i,\theta)]^{1-\delta_i}]$$

Therefore,

$$(2)\ LogL(\theta) = l(\theta) = \sum_{i=1}^{n}\delta_i \log[h(t_i,\theta)S(t_i,\theta)] + \sum_{i=1}^{n}(1-\delta_i)\log[S(t_i,\theta)]$$

$$= \sum_{i=1}^{n} \delta_i (\log[h(t_i, \theta)] + \log[S(t_i, \theta)]) + \sum_{i=1}^{n} (1 - \delta_i) \log[S(t_i, \theta)]$$

Because $\log[S(t)] = -H(t)$

$$(3) \quad l(\theta) = \sum_{i=1}^{n} \delta_i (\log[h(t_i, \theta)] - H(t_i, \theta)) - \sum_{i=1}^{n} (1 - \delta_i) \log[H(t_i, \theta)]$$

For a given data set, the value of θ that maximizes the log likelihood function is the MLE. Once I obtain the MLEs, I will plug them into the survival function or the cumulative hazard function and produce graphical displays to visualize the fit of the distribution to the data.

Residual Analysis

The Cox-Snell residuals are one way to investigate how well a model fits the data. They're calculated using the cumulative hazard function. Let T be continuous. Since the survival function is distributed uniformly on (0,1), i.e.,

$$S(T) \sim U(0,1)$$

it is not hard to show that the cumulative hazard function is exponentially distributed with θ=1,i.e.,

$$-\log S(T) = H(T) \sim \exp(\theta = 1),$$

The Cox-Snell residual, $\hat{r}_i$ is defined for the ith observation:

$$\hat{r}_i = \begin{cases} \hat{H}(t_i) & if\ the\ obeservation\ is\ a\ failure \\ \hat{H}(t_i) + 1 & if\ the\ observation\ is\ censored \end{cases}$$

I will generate a qqplot of the Cox-Snell Residuals versus an exponential distribution with mean 1. If the specified survival model fits the data well, I would expect to see the points align with a line with slope 1 and intercept 0. Because none of the distributions will be a perfect fit, one should look for the points to appear to be tightly scattered around the line to have good evidence that the model is a good fit to the data (Smith, 157-159).

Likelihood Ratio Test

One way to compare two nested models is by performing a likelihood ratio test. The test calculates how much more likely the data are if they came from one distribution compared to another. It compares the values of the likelihood function evaluated at the MLEs for each of the models. Since it's only possible to compare two models at a time, the simpler model is the null and the more complex model is the alternative. The ratio between the two models, or λ, is $\frac{L_o}{L_1}$ where L_o is the value of the likelihood function for the null model and L_1 is the value of the likelihood function for the alternative model. The test statistic for the likelihood ratio test, D, is negative twice the difference between the values of the log likelihood functions:

$$D = -2\,[\log(\lambda)]$$

$$D = -2\,[\log\left(\frac{L_o}{L_1}\right)]$$

$$D = -2[\log(L_o) - \log(L_1)]$$

where D follows a Chi-Square distribution ($D \sim (\chi^2_{df})$) and the degrees of freedom are equal to the number of free parameters in the alternative minus the number of free parameters in the null. If D is small, we conclude that the data are more likely to be from the null hypothesis model. If D is large, we conclude that the data are more likely to be from the more complex model ("Likelihood ratio tests").

AIC (Akaike Information Criteria)

A method for comparing among two or more models is the Akaike Information Criterion (AIC) which is a measure of the relative goodness of fit for statistical model. The AIC is calculated as negative two times the value of the log likelihood function plus two times the number of parameters, or

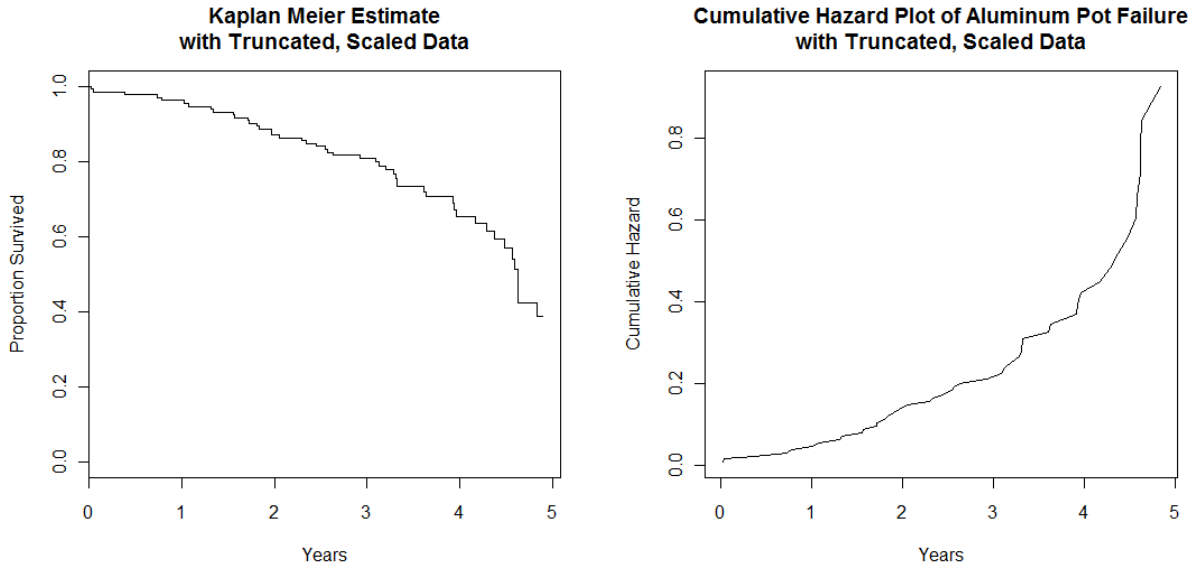
$$AIC = -2L(\hat{\theta}) + 2p.$$

A more complex model is usually a better predictive model, because more parameters allows for more flexibility in how the model fits the data (O'Meara). However, there can be a large increase in the complexity of a model while receiving only a little more predictive power. Therefore, the AIC penalizes the value of the log likelihood function for the complexity of the model, ensuring that the most complex model won't always be considered the best model ("Akaike's Information Criteria"). To use this method to choose among the best models, I will calculate the AIC value for each of the models, and then rank the models by this criterion. The model with the lowest AIC value is considered the best.

To summarize, it is important to use the visual representations as well as formal statistical tests to decide which model is the best. The visual representations, e.g. the Kaplan Meier curve, the cumulative hazard function plot, and the Cox-Snell residual plots, help make sure that the specified models actually fit the data. The likelihood ratio test and AIC value only measure which of the presented models is the best; they do not say if any of the models are even a good fit to the data. Once it is verified that a models fits the data by looking at the survival plot, cumulative hazard plot, and the residual plot, the likelihood ratio test and AIC value will be used to decide which model is the best for the data.

Chapter 3: Weibull and Gompertz Distributions

Recall that the Kaplan-Meier plot of the survival function and the cumulative hazard function for the aluminum pot data are



As seen in the cumulative hazard plot, the risk of failure is increasing nonlinearly over time. I will investigate 4 models that allow for increasing hazard rates. I will start with the most simple, and end with the most complex.

Weibull Distribution

The survival function of the Weibull distribution is

$$S(t) = e^{-(\theta t)^{\gamma}}, \text{ where } \theta > 0, \gamma > 0$$

and the pdf is

$$f(t) = \gamma \theta^{\gamma} t^{\gamma - 1} e^{-(\theta t)^{\gamma}}.$$

The hazard function is

$$h(t) = \gamma\theta^{\gamma}t^{\gamma-1}$$

and the cumulative hazard function is

$$H(t) = (\theta t)^{\gamma}.$$

The value of γ determines the direction of the failure rate. If $\gamma < 1$, the hazard function is decreasing, if $\gamma > 1$, the hazard function is increasing, and if $\gamma = 1$, the hazard function is constant. Note, when $\gamma = 1$

$$f(t) = 1\theta^{1}t^{1-1}e^{-(\theta t)^{1}} = \theta e^{-(\theta t)}$$

which is the exponential distribution with mean $\frac{1}{\theta}$. The exponential is a special case of the Weibull distribution ("Weibull distribution").

The log likelihood function of θ and γ is:

$$LogL(\theta,\gamma) = l(\theta,\gamma) = \sum_{i=1}^{n}\delta_i \log[f(t_i)] + \sum_{i=1}^{n}(1-\delta_i)\log[S(t_i)]$$

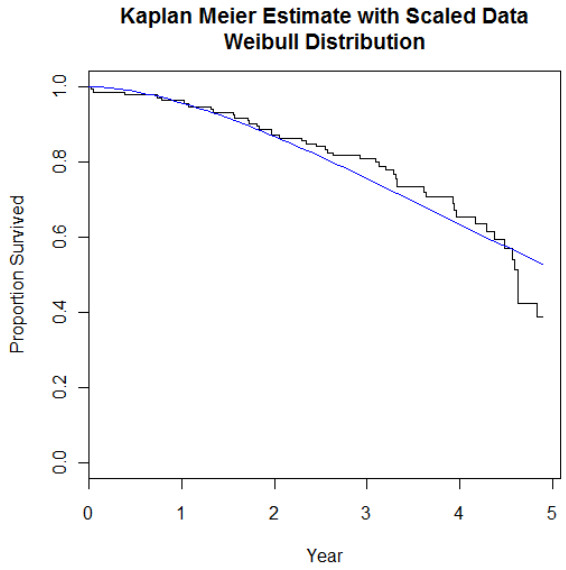
$$l(\theta,\gamma) = \sum_{i=1}^{n}\delta_i \log\left[\gamma\theta^{\gamma}t^{\gamma-1}e^{-(\theta t)^{\gamma}}\right] + \sum_{i=1}^{n}(1-\delta_i)\log\left[e^{-(\theta t)^{\gamma}}\right]$$

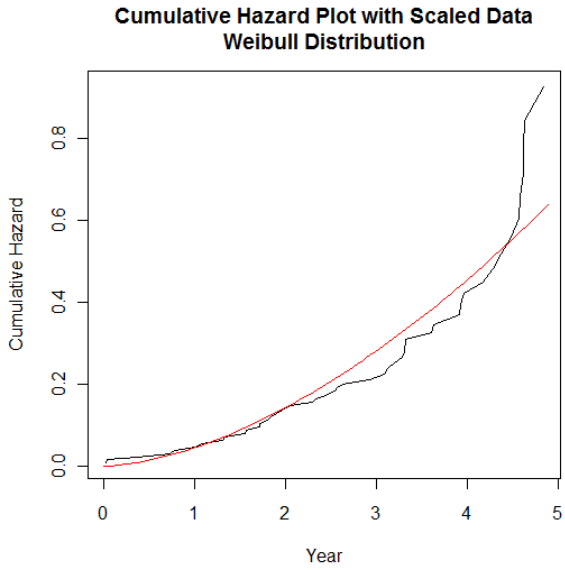
$$l(\theta,\gamma) = \sum_{i=1}^{n}\delta_i[\log(\gamma\theta^{\gamma}) + (\gamma-1)\log(t_i) - (\theta t_i)^{\gamma}] + \sum_{i=1}^{n}(1-\delta_i)[-(\theta t_i)^{\gamma}]$$

The maximization of the log likelihood function requires numerical methods. To do this, I wrote create a function in R that returns the value of the negative log likelihood function. I then use an optimization function that will minimize this function to obtain the maximum likelihood estimates of θ and γ (Steenbergen,2-6). Refer to section 3 of the Appendix for the R code.

Results

For the aluminum pot data, the maximum likelihood estimates are: $\hat{\theta}$=0.156 and $\hat{\gamma}$=1.680. The standard errors of the estimates found by inverting the Hessian are: $se(\hat{\theta})$=0.018 and $se(\hat{\gamma})$=0.227. To see how well the Weibull distribution fits these data, I plug the MLEs into the survival function and plot the Weibull survival curve over the Kaplan-Meier estimate of the survival curve. I also plot the MLE of the cumulative hazard function on the Kaplan-Meier estimate of the cumulative hazard function. See the figures below. As can be seen from the plots of the survival function and the cumulative hazard function, the Weibull distribution is a little too rigid of a distribution to fit these data. It appears this distribution fits well from time 0 to about 2 years, but not so well after that. Both graphs show that the Weibull may be a little better if it was more flexible or there was more curvature after time of 3 years.





Gompertz Distribution

The survival function of the Gompertz distribution is

$$S(t) = e^{-\frac{\beta}{\gamma}(e^{\gamma t}-1)}, \text{ where } \beta > 0, \gamma > 0$$

and the pdf is

$$f(t) = \beta e^{\gamma t - \frac{\beta}{\gamma}(e^{\gamma t}-1)}.$$

The hazard function is

$$h(t) = \beta e^{\gamma t}$$

and the cumulative hazard function is

$$H(t) = \frac{\beta}{\gamma}(e^{\gamma t} - 1).$$

The log likelihood function is constructed as follows (Hogg and Ledolter, 120):

$$l(\beta, \gamma) = \sum_{i=1}^{n} \delta_i \log[f(t_i)] + \sum_{i=1}^{n} (1 - \delta_i) \log[S(t_i)]$$

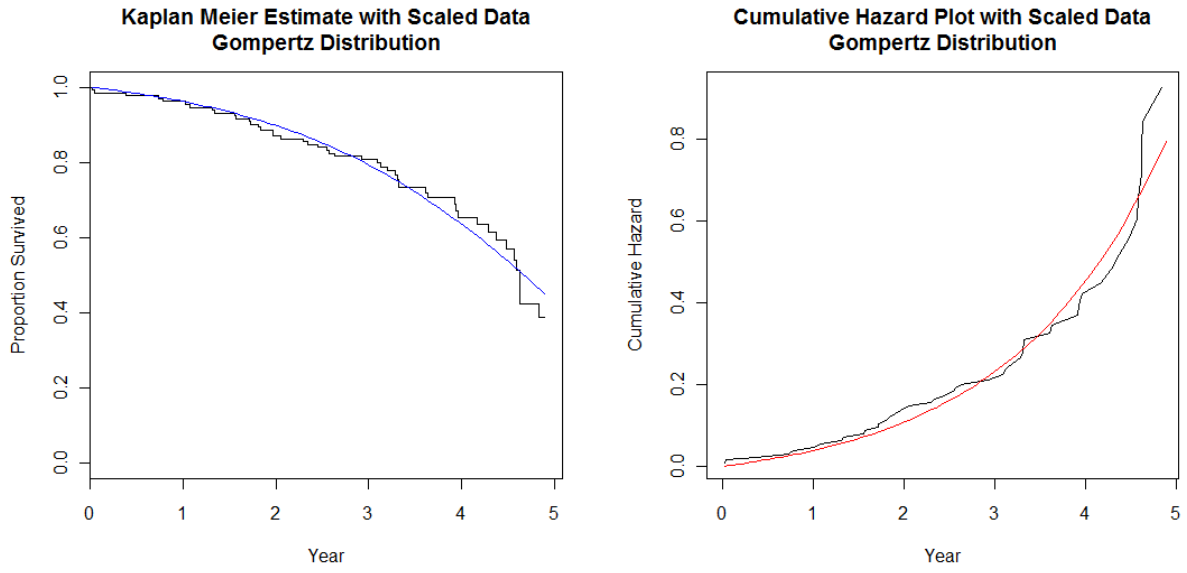
$$l(\beta, \gamma) = \sum_{i=1}^{n} \delta_i \log\left[\beta e^{\gamma t - \frac{\beta}{\gamma}(e^{\gamma t}-1)}\right] + \sum_{i=1}^{n} (1 - \delta_i) \log\left[e^{-\frac{\beta}{\gamma}(e^{\gamma t}-1)}\right]$$

$$l(\beta, \gamma) = \sum_{i=1}^{n} \delta_i \left[\log(\beta) + \gamma t_i + \left(-\frac{\beta}{\gamma}(e^{\gamma t_i - 1})\right)\right] + \sum_{i=1}^{n} (1 - \delta_i)\left[-\frac{\beta}{\gamma}(e^{\gamma t_i} - 1)\right]$$

The maximization of the log likelihood function requires numerical methods. To do this, I wrote another function in R that returns the negative log likelihood function of this distribution. I then use an optimization function that will minimize the function to obtain the maximum values of β and γ (Steenbergen,2-6). Refer to section 4 of the Appendix for the R code.

Results

The maximum likelihood estimates are: $\hat{\beta}$ =0.028 and $\hat{\gamma}$ =0.585. The standard errors of the estimates are: $se(\hat{\beta})$= 0.010 and $se(\hat{\gamma})$=0.121. To see how well the Gompertz distribution fits these data, I plug the MLEs into the survival function and plot the Gompertz survival curve over the Kaplan-Meier estimate. See the figures below. As can be seen from the plots of the survival function and cumulative hazard function that, the fit of the Gompertz distribution to those data appears to be more flexible than the Weibull and appears to model the exponentially increasing failure rates of the data better. From the cumulative hazard function, plot it appears that this model does a good job fitting the data until about time 3.5 years, but not such a good job in the upper tail.



Both Weibull and the Gompertz distributions are two parameter models. In the next two chapters, I consider more complex models to try to improve the fit to the pot data.

Chapter 4: Gompertz-Makeham Distribution

The Gompertz-Makeham law states that the failure rate is a combination of an age independent Makeham term and an age dependant Gompertz term. The failure rate of the Makeham term is linear, while the failure rate of the Gompertz term is exponential. This law describes the behavior of human mortality, most accurately in the later part of life ("Gompertz-Makeham law of mortality"). The Gompertz-Makeham distribution is an extension of the Gompertz distribution and a more complex distribution. Recall the survival function of the Gompertz distribution is

$$S(t) = e^{-\frac{\beta}{\gamma}(e^{\gamma t}-1)}.$$

The survival function of the Gompertz-Makeham distribution is

$$S(t) = e^{-\alpha t-\frac{\beta}{\gamma}(e^{\gamma t}-1)} = e^{-\alpha t}e^{-\frac{\beta}{\gamma}(e^{\gamma t}-1)}$$

The Gompertz-Makeham distribution differs from the Gompertz because of the additional term, $e^{-\alpha t}$. Looking at the cumulative hazard function of the Gompertz-Makeham distribution,

$$H(t) = \alpha t + \frac{\beta}{\gamma}(e^{\gamma t} - 1)$$

this addition allows the hazard rate to change over time. For small values of t we see a more linear increase in the risk of failure over time. For large values of t the exponential term will dominate ($\frac{\beta}{\gamma}(e^{\gamma t} - 1)$) and we will see an exponential increase in the failure rate. From the Kaplan-Meier plot of the survival function and the cumulative hazard function, it appears that times 0 to about 3 years have a more shallow slope than times greater than 3. This distribution was chosen as a possible fit to the data because the extra term in the Gompertz-Makeham distribution may do a better job fitting the more linear part of the underlying function while the Gompertz term may fit the data with more curvature.

Gompertz-Makeham Distribution

As stated, the survival function of the Gompertz-Makeham distribution is

$$S(t) = e^{-\alpha t - \frac{\beta}{\gamma}(e^{\gamma t}-1)}, \text{ where } \alpha > 0, \beta > 0, \gamma > 0.$$

The pdf for this distribution is

$$f(t) = (\alpha + \beta e^{\gamma t}) e^{-\alpha t - \frac{\beta}{\gamma}(e^{\gamma t}-1)}.$$

The hazard function is

$$h(t) = \alpha + \beta e^{\gamma t}$$

and the cumulative hazard function is

$$H(t) = \alpha t + \frac{\beta}{\gamma}(e^{\gamma t} - 1).$$

The log likelihood function of α, β, and γ is

$$l(\alpha, \beta, \gamma) = \sum_{i=1}^{n} \delta_i \log[f(t_i)] + \sum_{i=1}^{n} (1 - \delta_i) \log[S(t_i)]$$

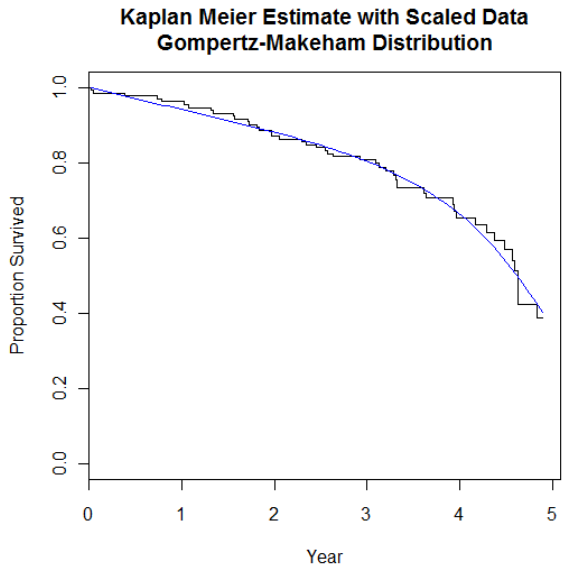
$$l(\alpha, \beta, \gamma) = \sum_{i=1}^{n} \delta_i \log\left[(\alpha + \beta e^{\gamma t}) e^{-\alpha t - \frac{\beta}{\gamma}(e^{\gamma t}-1)}\right] + \sum_{i=1}^{n} (1 - \delta_i) \log\left[e^{-\alpha t - \frac{\beta}{\gamma}(e^{\gamma t}-1)}\right]$$

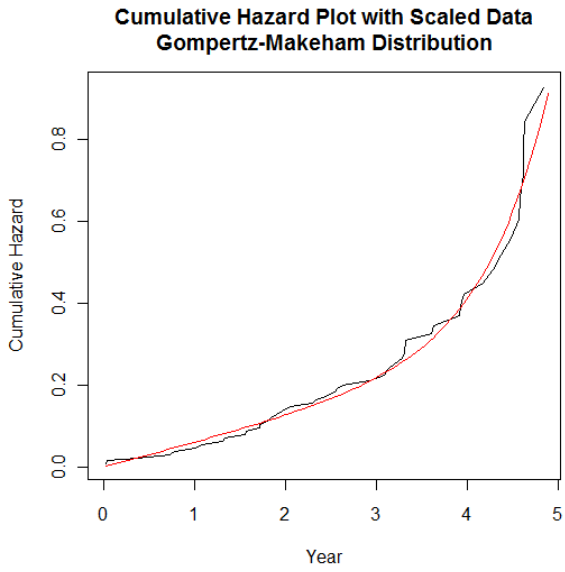
$$l(\alpha, \beta, \gamma) = \sum_{i=1}^{n} \delta_i \left[\log(\alpha + \beta e^{\gamma t_i}) + \left(-\alpha t_i - \frac{\beta}{\gamma}(e^{\gamma t_i - 1})\right)\right] + \sum_{i=1}^{n} (1 - \delta_i) \left[-\alpha t_i - \frac{\beta}{\gamma}(e^{\gamma t_i} - 1)\right]$$

The maximization of the log likelihood function requires numerical methods. To do this, I wrote create a function in R that returns the value of the negative log likelihood function. I then use an optimization function that will minimize this function to obtain the maximum likelihood of α, β and γ (Steenbergen,2-6). Refer to section 5 of the Appendix for the R code.

Results

The maximum likelihood estimates are: $\hat{\alpha}$=0.058, $\hat{\beta}$=0.0009, and $\hat{\gamma}$=1.4111. I was not able to obtain the standard errors for these estimates. To see how well the Gompertz-Makeham distribution fits these data, I plug the MLEs into the survival function and plot the Gompertz-Makeham survival curve over the Kaplan-Meier estimate of the survival curve. I also plot the MLE of the cumulative hazard function over the Kaplan-Meier estimate. See the figures below. It appears the Gompertz-Makeham distribution does a good job of fitting the distribution of pot failure times.





One reason to consider the Gompertz-Makehman distribution was that the distribution has the flexibility to fit both the early and the later failure times. In the next chapter I consider another model that also has this feature.

Chapter 5: Weibull Change Point Model

Inspection of the cumulative hazard plot for the pot data suggests that the hazard rate increases linearly until about year 3 (actually 6.4 years from the start of production). After that, the pots start failing more rapidly. In this chapter I propose a Weibull change point survival model to better model this change in the failure rates of the aluminum pots. The model specifies a Weibull distribution for the failure times prior to a specified time point which I called "a" and then specifies another Weibull distribution for the failure times after time "a".

Weibull Change Point Model

Recall the Weibull the survival function is

$$S(t) = e^{-(\theta t)^{\gamma}}$$

and that the cumulative hazard function is

$$H(t) = -\log S(t) = (\theta t)^{\gamma}.$$

Define $g(t)$ to be the logH(t), where

$$g(t) = \log H(t) = \alpha \log(\theta t) = \alpha \log(\theta) + \alpha \log(t) = \beta + \alpha \log(t)$$

and $\beta = \alpha \log(\theta)$.

Let the change point be denoted by "a" and define an indicator variable c such that

$$c = \begin{cases} 1 & if\ 0 \le t \le a \\ 0 & otherwise \end{cases}$$

The specify the Weibull change point model, I will define the log of the cumulative hazard function as:

$$g(t) = c[\beta_1 + \alpha_1 \log(t)] + (1 - c)[\beta_2 + \alpha_2 \log(t)]$$

where $g(t) = [\beta_1 + \alpha_1 \log(t)]$ if $t \le \alpha$ and $g(t) = [\beta_2 + \alpha_2 \log(t)]$ if $t > \alpha$. To ensure that the function is smooth at the change point "a", it is necessary that

$$\beta_1 + \alpha_1 \log(a) = \beta_2 + \alpha_2 \log(a)$$

Therefore,

$$\beta_2 = \beta_1 + (\alpha_1 - \alpha_2)\log(a).$$

By substituting $\beta_1 + (\alpha_1 - \alpha_2)\log(a)$ for β_2, it follows that

$$g(t) = c[\beta_1 + \alpha_1 \log(t)] + (1-c)[\beta_1 + (\alpha_1 - \alpha_2)\log(a) + \alpha_2 \log(t)]$$

$$g(t) = \beta_1 + c\alpha_1 \log(t) + (1-c)[(\alpha_1 - \alpha_2)\log(a) + \alpha_2 \log(t)]$$

Recall that $\beta_1 = \alpha_1 \log(\theta)$, therefore

$$\log H(t) = \alpha_1 \log(\theta) + c\alpha_1 \log(t) + (1-c)[(\alpha_1 - \alpha_2)\log(a) + \alpha_2 \log(t)]$$

This model has four parameters, a, α_1, α_2 and θ.

To find the survival function for this change point model I need to find the hazard function. Recall that

$$h(t) = \frac{d}{dt} H(t).$$

Since

$$g(t) = \log H(t), H(t) = e^{g(t)},$$

Hence by the chain rule,

$$h(t) = [e^{g(t)}] \frac{d(g(t))}{dt}$$

Taking the derivative of $g(t)$ with respect to t,

$$\frac{d(g(t))}{dt} = \frac{c\alpha_1}{t} + \frac{(1-c)\alpha_2}{t}$$

and therefore

$$h(t) = e^{g(t)}\left[\frac{c\alpha_1}{t} + \frac{(1-c)\alpha_2}{t}\right]$$

or

$$h(t) = H(t)\left[\frac{c\alpha_1}{t} + \frac{(1-c)\alpha_2}{t}\right] = \frac{H(t)}{t}[c\alpha_1 + (1-c)\alpha_2]$$

Finally, to find the pdf for the change point Weibull, recall

$$S(t) = e^{-H(t)}$$

and that

$$f(t) = h(t)e^{-H(t)}$$

The log-likelihood function for the Weibull change point model is

$$l(a, \alpha_1, \alpha_2, \theta) = \sum_{i=1}^{n} \delta_i \left[\mathrm{c_i}\log \alpha_1 + (1 - c_i)\log\ \alpha_2\right] - \delta_i log t_i + \delta_i log H(t_i) - H(t_i)$$

The four parameters of interest are, a, α_1, α_2andθ.

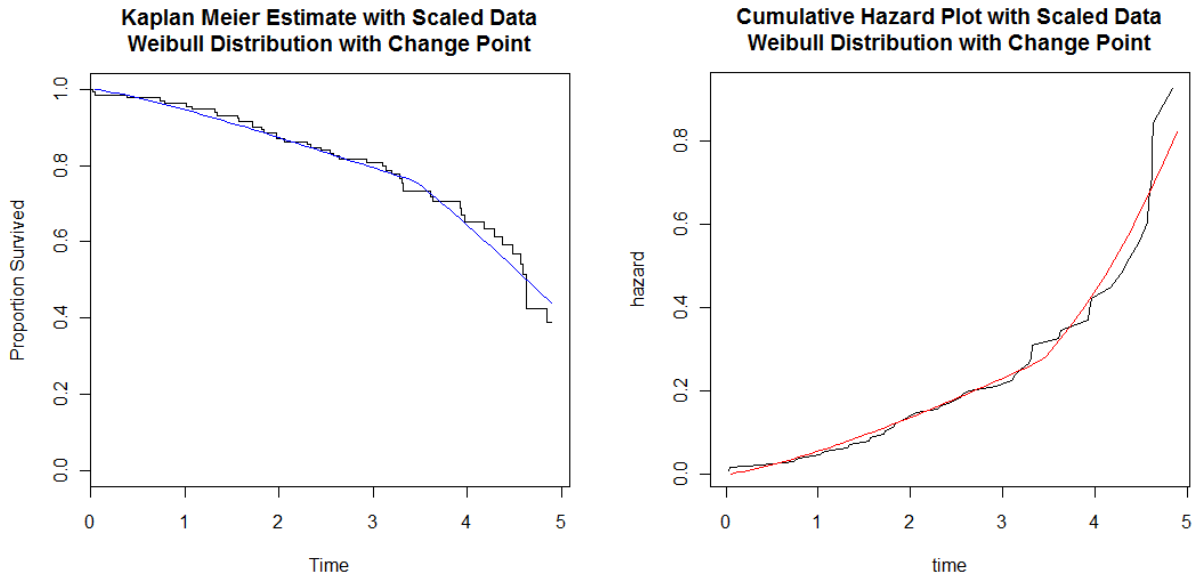
Finding the Change Point

To fit this model, it is necessary to find the change point "a". While it is possible to look at the cumulative hazard plot and select a value of a, it is also possible to find the change point formally using maximum likelihood estimation. To find the MLE of the change point, I use the following method. I choose a range of values of "a" around where I think the change point is. I choose the smallest value of "a" in this range and maximize the likelihood function for the other parameters in the model. I note the value of the log likelihood function at the maximum. Then I choose the next value of "a" in this range and repeat the procedure. I do this for all the values of a in the range. I then identify the value of "a" that gives the largest log likelihood function. This is the MLE of "a." This procedure generates the profile likelihood of "a" and yields the joint MLEs of all of the parameters.

Results

The maximum likelihood estimates are a=3.44, α_1=1.304, α_2=3.107, and θ=0.108. The standard errors of the estimates are $se(\widehat{\alpha_1}) = 0.224, se(\widehat{\alpha_2}) = 0.633$ and $se(\hat{\theta}) = 0.025$. Using these parameters, it's now possible to plot the survival function and the cumulative hazard

function; see plots below. The change point does a great job fitting the failure times before time 3.44 years and the fit after time 3.44 years is pretty good too.



I have fit four distributions to the aluminum pot data. The next step is to test which of the distributions fits the data the best using the likelihood ratio test and AIC.

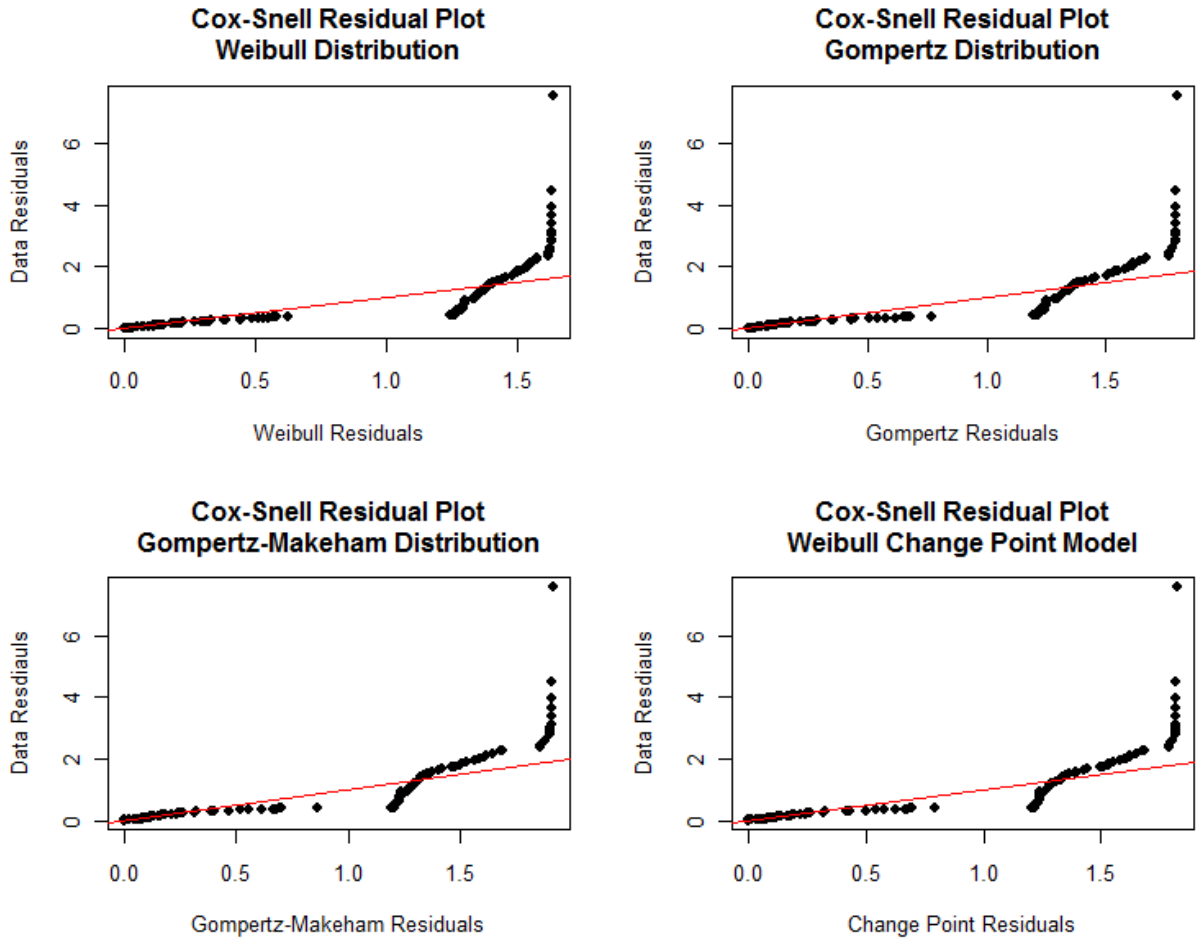
Chapter 6: Choosing Among Models

I have fit the four survival models to aluminum pot failure time data. The next step is to investigate differences in the fits of the different models.

Cox-Snell Residuals

First I will investigate the fit of the models using the using Cox-Snell residuals. For each of the survival models, I calculate the estimated cumulative hazard function for each pot by substituting the estimated MLEs for the parameters of the respective model. Recall from Chapter 2, if the pot is a failure, the Cox-Snell residual is equal to the estimated value of the hazard function, and if the pot is a censored observation, the Cox-Snell residual is equal to the estimate of the calculated value of the cumulative hazard function plus 1.

The Cox-Snell residual plots for all four models are on the following page. As was pointed out in Chapter 2, if a model fits well, we expect the Cox-Snell residuals to follow an exponential distribution with mean=1. The plots on the next page are qqplots. In the plots on the next page the red line slope=1 and intercept=0. If the model fits well we would expect the residuals to fall along this line. For all four models we see that the tail on the right hand side deviates from the red line. The aberrations we see in the Cox-Snell residual plots below emphasize where the models don't fit the data well. Perhaps it is not surprising that none of the models fit well in the upper tail where there are only censored observations and no failures. In addition to assessing the model fits visually, it is also important to perform formal statistical test, such as calculate the AIC and perform the log likelihood ratio test.



AIC

After calculating the value of the log likelihood function evaluated at the MLEs, the AIC value is obtained by :

$$AIC = -2\hat{l}(\theta) + 2p$$

where $\hat{l}(\theta)$ is the value of the log likelihood equation function evaluated at the MLEs and p is the number of parameters in the model. The AIC is calculated for each survival model. The model with the smallest AIC is the model with the best fit among the models being considered. The results are summarized in the table below

	Weibull	Gompertz	Change Point	Gompertz-Makeham
Log Likelihood Value	-145.90	-140.04	-141.63	-138.64
Number of Parameters	2	2	4	3
AIC	295.79	284.08	291.26	283.28
Rank	**4**	**2**	**3**	**1**

As we suspected from inspection of the fitted survival curves, cumulative hazard plots, and Cox-Snell residual plots, the Gompertz-Makeham distribution is the best fit for the failure time distribution of the aluminum pot data. This is probably because it is the most flexible model which allows for a small and constant failure rate for young pots and a larger and exponentially increasing failure rate for older pots. However, the AIC value for the Gompertz distribution is just a little larger than the AIC value for the Gompertz-Makeham distribution.

Likelihood Ratio Test

The comparison of the nested models can be done using the likelihood ratio test. The likelihood ratio test statistics is given by:

$$D = -2[\log(L_o) - \log(L_1)].$$

D follows a Chi-Square distribution, with degrees of freedom equal to the number of free parameters in the alternative distribution minus the number of parameters in the null distribution. Because the Weibull and Weibull change point survival models are nested, we can compare them using the likelihood ratio test. Specifically, the Weibull change point model has four parameters a, α_1, α_2 and θ, and the smaller model, i.e., the Weibull distribution, is obtained when $\alpha_1 = \alpha_2 = \alpha$. When $\alpha_1 = \alpha_2 = \alpha$,

$$H(t) = e^{\alpha_1 \log(\theta) + c\alpha_1 \log(t) + (1-c)[(\alpha_1+\alpha_2)\log(a) + \alpha_2 \log(t)]}$$

$$H(t) = e^{\log(\theta)^\alpha + \log(t)^\alpha} = e^{\log(\theta)^\alpha} e^{\log(t)^\alpha} = \theta t^\alpha.$$

This is the cumulative hazard function of the Weibull distribution. Therefore, the null hypothesis is $\alpha_1 = \alpha_2$. Using the table above, and the equation for the likelihood ratio test:

$$D = -2[\log(L_o) - \log(L_1)]$$

$$D = -2[(-145.90) - (-141.63)] = 8.54 \; with \; 4 - 2 = 2 \; degrees \; of \; freedom$$

The critical value of a Chi Squared distribution with 2 degrees of freedom at the 0.05 level is 5.99. Since the test statistic is 8.54, and 8.54>5.99, the likelihood ratio test is significant at the 0.05 level, which means we reject the null hypothesis that $\alpha_1 = \alpha_2$. We conclude the change point model is a better fit than a single Weibull, which is consistent with the analysis based on the AIC value.

A similar procedure is used when comparing the Gompertz and the Gompertz-Makeham distributions. The Gompertz distribution has two parameters, $\gamma$and β, and its cumulative hazard function is

$$H(t) = \frac{\beta}{\gamma}(e^{\gamma t} - 1)$$

while the Gompertz-Makeham has three parameters, α, γ and β, and its cumulative hazard function is

$$H(t) = \alpha t + \frac{\beta}{\gamma}(e^{\gamma t} - 1).$$

It's quite simple to see that when $\alpha = 0$, the Gompertz-Makeham distribution becomes the Gompertz distribution. Therefore, the null hypothesis is $\alpha = 0$. Using the values of the log likelihood function from the table above,

$$D = -2[\log(L_o) - \log(L_1)]$$

$$D = -2[(-140.0389) - (-138.6386)] = 2.80 \; with \; 3 - 2 = 1 \; degrees \; of \; freedom.$$

The critical value of a Chi Squared distribution with 1 degree of freedom at the 0.05 level is 3.84. Since our test statistic is 2.80, and 2.80<3.84, the likelihood ratio test is not significant at the 0.05 level.. We cannot reject the null hypothesis that $\alpha = 0$and conclude that the Gompertz-Makeham distribution is not a significantly better fit to our data than the Gompertz. Because the AIC values for the Gompertz and Gompertz-Makeham are so close, this is not a surprise.

Chapter 7: Discussion

Using the AIC values to compare all four distributions, it is found that the Gompertz-Makeham distribution is considered the best fit for the distribution of the failure times for the pot data. However, the likelihood ratio test indicates that the Gompertz-Makehm model does not provide a significantly better fit than the Gompertz distribution, which was ranked second when looking at the AIC values. Since the AIC value for the Gompertz model is smaller than the Weibull change point model, I conclude that the best model for the data is the Gompertz survival distribution.

The results of this thesis suggest directions for further analysis. There were various covariates collected during the observation of these pots, such as average temperature, average voltage, number of times the ratio in the bath dropped below threshold, etc. Now that I have identified a survival model that best fits the data, future work would be to develop a Gompertz survival model that includes covariates to more accurately estimate when a pot will fail.

While I have fit a Gompertz distribution here, I have fit it to the transformed version of the data, which has been truncated and scaled. Recall that in the original data, there was a long period without any failures, which is why the data were truncated. An extension of the Gompertz survival model could include threshold parameters to account for this early period without failures.

From investigating this problem and doing this research, I have gained a significant portion of understanding of survival analysis. I learned the importance of the different representations of the survival distribution and how they give different insights into the data. The importance of censored observations has also become very clear, as well as how to make sure they're accounted for correctly in the data. I have also learned how to think through

problems that come up during analyses and how to look for alternative solutions when the problem cannot be solved. The deep understanding of the structure of statistical models and methods is an integral part of any analysis.

Ultimately, in order to fit a probability distribution to a data set, it is important to first get an understanding of the structure of the data. This will help in the specification of a model. After choosing possible models, use maximum likelihood estimation to estimate the MLEs for each distribution, and use them to visualize the survival and hazard functions to get an idea of the fit. Then, perform formal tests, such as computing the AIC and the likelihood ratio test to choose which model fits the best to the data.

Appendix: R Code

Section 1

Chapter 1: Light bulb example

```
library(survival)
lightbulb=rexp(30,1)
lb.ft=survfit(Surv(lightbulb)~1)
plot(lb.ft, ylab="Survival Function",xlab="Days (y)", main="Kaplan Meier Curve of Light Bulb
        Lifetime", mark.time=F,conf.int=F)
lb.haz=coxph(Surv(lightbulb)~1)
plot(basehaz(lb.haz)[,2:1], type="l", main="Cumulative Hazard Plot of Light Bulb Lifetime")
exp.surv=function(theta,y){
        surv=theta*exp(-theta*y)
        return(surv)
        }
plot(lb.ft, main="Kaplan Meier Curve of Data From an\nExponential Distribution",
        conf.int=F,mark.time=F)
curve(exp.surv(1,x), 0, max(lightbulb),col="blue",add=T)
exp.haz=function(theta,y){
haz=(theta*y)
return(haz)
}
plot(basehaz(lb.haz)[,2:1], type="l", main="Cumulative Hazard Plot\nExponential Distribution")
curve(exp.haz(1,x), 0, max(lightbulb),col="red",add=T)
```

Section 2

Kaplan-Meier estimate for the original data set

```
surv.age.fit=survfit(Surv(Age, status01)~1)
plot(surv.age.fit, main="Kaplan Meier Estimate of Failure Time", xlab="Days",
        ylab="Proportion Survived", mark.time=F, conf.int=F)
pot.age=coxph(Surv(Age, status01)~1)
plot(basehaz(pot.age)[,2:1], type="l", main="Cumulative Hazard Plot of Aluminum Pot Failure",
        xlab="Days",ylab="Cumulative Hazard")
```

Kaplan-Meier estimate for the conditional data set

```
surv.fit=survfit(Surv(t, status01)~1)
plot(surv.fit, main="Kaplan Meier Estimate with Truncated Data", xlab="Days",
        ylab="Proportion Survived", mark.time=F, conf.int=F)
pot.cox=coxph(Surv(t, status01)~1)
plot(basehaz(pot.cox)[,2:1], type="l", main="Cumulative Hazard Plot of Aluminum Pot
        Failure\nwith Truncated Data", xlab="Days",ylab="Cumulative Hazard")
```

Kaplan-Meier estimate for the scaled data set

```
surv.data.scale=Surv(t.scale, status01)~1
surv.fit.scale=survfit(surv.data.scale)
pot.cox.scale=coxph(Surv(t.scale, status01)~1)
plot(basehaz(pot.cox.scale)[,2:1], type="l", main="Cumulative Hazard Plot of Aluminum Pot
        Failure\nwith Truncated, Scaled Data",xlab="Years",ylab="Cumulative Hazard")
```

Section 3

Log likelihood function for the Weibull model

```
weib.likl<-function(param,y){
theta<-param[1]
        gamma<-param[2]
        logl<-sum(y[,2]*(log(gamma) + gamma*log(theta) + (gamma-1)*log(y[,1]) -
                (theta*y[,1])^gamma )) -sum((1-y[,2])*(theta*y[,1])^gamma)
        return(-logl)
}
param.weib=optim(c(0.1,0.1),weib.likl,y=data.scale,hessian=T)$par
p.weib=optim(c(0.1,0.1),weib.likl,y=data.scale,hessian=T)
OI.weib=solve(p.weib$hessian)
se.weib=sqrt(diag(OI.weib))
```

Plotting the survival function for the Weibull model

```
weib.surv<-function(param,y){
theta<-param[1]
gamma<-param[2]
survival<-exp(-(theta*y)^gamma)
return(survival)
}
plot(surv.fit.scale, main="Kaplan Meier Estimate with Scaled Data\nWeibull Distribution",
        xlab="Time", ylab="Proportion Survived", mark.time=F, conf.int=F)
curve(weib.surv(param.weib,x),0,max(t.scale), col="blue", add=T)
```

Plotting the hazard function for the Weibull model

```
weib.haz<-function(param,y){
theta<-param[1]
gamma<-param[2]
haz<--log(weib.surv(param,y))
return(haz)
}
plot(basehaz(pot.cox.scale)[,2:1], type="l", main="Cumulative Hazard Plot with Scaled
        Data\nWeibull Distribution", xlab="Years",ylab="Cumulative Hazard"))
curve(weib.haz(param.weib,x),0,max(t.scale), col="red", add=T)
```

Section 4

Log likelihood function for the Gompertz model

```
gomp.likl<-function(param,y){
        beta<-param[1]
        gamma<-param[2]
        logl<-sum(y[,2]*(log(beta)+gamma*y[,1]+(-(beta/gamma)*(exp(gamma*y[,1])-1)))) +
        sum((1-y[,2])*(-(beta/gamma)*(exp(gamma*y[,1])-1)))
        return(-logl)
}
param.gomp<-optim(c(.2,.2),gomp.likl,y=data.scale)$par
p.gomp<-optim(c(.2,.2),gomp.likl,y=data.scale,hessian=T)
OI.gomp=solve(p.gomp$hessian)
se.gomp=sqrt(diag(OI.gomp))
```

Plotting the survival function for the Gompertz model

```
gomp.surv<-function(param,y){
        beta<-param[1]
        gamma<-param[2]
        surv<-exp(-(beta/gamma)*(exp(gamma*y)-1))
        return(surv)
}
plot(surv.fit.scale, main="Kaplan Meier Estimate with Scaled Data\nGompertz Distribution",
        xlab="Time", ylab="Proportion Survived",mark.time=F, conf.int=F)
curve(gomp.surv(param.gomp,x),min(t.scale),max(t.scale), col="blue",add=T)
```

Plotting the hazard function for the Gompertz model

```
gomp.haz<-function(param,y){
        beta<-param[1]
        gamma<-param[2]
        haz<--log(gomp.surv(param,y))
        return(haz)
}
plot(basehaz(pot.cox.scale)[,2:1], type="l", main="Cumulative Hazard Plot with Scaled
        Data\nGompertz Distribution", xlab="Years",ylab="Cumulative Hazard")))
curve(gomp.haz(param.gomp,x),min(t.scale),max(t.scale),col="red",add=T)
```

Section 5

Log likelihood function for the Gompertz-Makeham model

```
gompmak.likl<-function(param,y){
        beta<-param[1]
        gamma<-param[2]
        alpha<-param[3]
```

```
logl<-sum(y[,2]*(log(alpha+beta*exp(gamma*y[,1])) + (-alpha*y[,1]-
        (beta/gamma)*(exp(gamma*y[,1])-1)))) + sum((1-y[,2])*(-alpha*y[,1]-
        (beta/gamma)*(exp(gamma*y[,1])-1)))
return(-logl)
}
param.gm=optim(c(.03,0.6,.11),gompmak.likl,y=data.scale)$par
p.gm=optim(c(.03,0.11,.6),gompmak.likl,y=data.scale,hessian=T)
optim(c(.01,.09,.45),gompmak.likl,y=data.scale)$par
OI.gm=solve(p.gm$hessian)
se.gm=sqrt(diag(OI.gm))
```

Plotting the survival function of the Gompertz-Makeham model

```
gm.surv<-function(param,y){
beta<-param[1]
gamma<-param[2]
alpha<-param[3]
surv<-exp(-alpha*y-(beta/gamma)*(exp(gamma*y)-1))
return(surv)
}
plot(surv.fit.scale, main="Kaplan Meier Estimate with Scaled Data\nGompertz-Makeham
        Distribution", xlab="Time", ylab="Proportion Survived", mark.time=F, conf.int=F)
curve(gm.surv(param.gm,x),min(t.scale),max(t.scale), col="blue",add=T)
```

Plotting the hazard function of the Gompertz-Makeham mode.

```
gm.haz<-function(param,y){
        beta<-param[1]
        gamma<-param[2]
        alpha<-param[3]
        haz<--log(gm.surv(param,y))
        return(haz)
}
plot(basehaz(pot.cox.scale)[,2:1], type="l", main="Cumulative Hazard Plot with Scaled
        Data\nGompertz-Makeham Distribution", xlab="Years",ylab="Cumulative Hazard")))
curve(gm.haz(param.gm,x),min(t.scale),max(t.scale), col="red",add=T)
```

Section 6

Choosing the change point for the Weibull change point model

```
a=seq(3,3.5, by=.01)
value.vec=rep(0,length(a))
weib.likl.a<-function(param,y,a){
        alpha1<-param[1]
        alpha2<-param[2]
```

```
        theta<-param[3]
        logl<-sum(y[,2]*(y[,3]*log(alpha1) + (1-y[,3])*log(alpha2)) -
                y[,2]*log(y[,1])+y[,2]*(alpha1*log(theta) + y[,3]*alpha1*log(y[,1]) + (1-
                y[,3])*((alpha1-alpha2)*log(a) + alpha2*log(y[,1]))) - exp(alpha1*log(theta) +
                y[,3]*alpha1*log(y[,1]) + (1-y[,3])*((alpha1-alpha2)*log(a) +
                alpha2*log(y[,1])))
        return(-logl)
}
for (i in 1:length(a)){
        c=rep(0,length(t))
        c[t<=a[i]]=1
        data.cp=cbind(t,status01,c)
        value.vec[i]=optim(c(1.34,2.81,0.11), weib.likl.a, y=data.cp,a=a[i])$value
}
a[which(value.vec==max(value.vec))]
```

Log likelihood function of the Weibull change point model

```
c.scale=rep(0,length(t.scale))
c.scale[t.scale<=(3.44)]=1
data.cp.scale=cbind(t.scale,status01,c.scale)
weib.likl.cp<-function(param,y,a){
        alpha1<-param[1]
        alpha2<-param[2]
        theta<-param[3]
        logl<-sum(y[,2]*(y[,3]*log(alpha1) + (1-y[,3])*log(alpha2)) - y[,2]*log(y[,1]) +
                y[,2]*(alpha1*log(theta) + y[,3]*alpha1*log(y[,1]) + (1-y[,3])*((alpha1-
                alpha2)*log(a) + alpha2*log(y[,1]))) - exp(alpha1*log(theta) +
                y[,3]*alpha1*log(y[,1]) + (1-y[,3])*((alpha1-alpha2)*log(a) +
                alpha2*log(y[,1])))
        return(-logl)
}
param.cp.scale<-optim(c(1.34,2.81,0.11), a=(3.44), weib.likl.cp, y=data.cp.scale,hessian=T)$par
p.cp<-optim(c(1.34,2.81,0.11), a=(3.44), weib.likl.cp, y=data.cp.scale,hessian=T)
OI.cp=solve(p.cp$hessian)
se=sqrt(diag(OI.cp))
```

Plotting the survival function of the Weibull change point model

```
weib.cp.surv<-function(param,y){
        alpha1<-param[1]
        alpha2<-param[2]
        theta<-param[3]
        c=rep(0,length(y))
```

```
        c[y<=(3.44)]=1
        a=3.44
        survival<-exp(-exp(alpha1*log(theta) + c*alpha1*log(y) +(1-c)*((alpha1-alpha2)*log(a)
                + alpha2*log(y))))
        return(survival)
}
plot(surv.fit.scale, main="Kaplan Meier Estimate with Scaled Data\nWeibull Distribution with
        Change Point", xlab="Time", ylab="Proportion Survived", mark.time=F, conf.int=F)
curve(weib.cp.surv(param.cp.scale,x),0,max(t.scale),col="blue", add=T)
```

Plotting the hazard function of the Weibull change point model

```
weib.cp.haz<-function(param,y){
        alpha1<-param[1]
        alpha2<-param[2]
        theta<-param[3]
        a=3.44
        haz=-log(weib.cp.surv(param.cp.scale,y))
        return(haz)
}
plot(basehaz(pot.cox.scale)[,2:1], type="l", main="Cumulative Hazard Plot with Scaled
        Data\nWeibull Distribution with Change Point", xlab="Years",ylab="Cumulative
        Hazard")))
curve(weib.cp.haz(param.cp.scale,x),0,max(t.scale), col="red", add=T)
```

Section 7

Cumulative hazard functions for all four models

```
cum.haz.weib=function(param,y){
        theta=param[1]
        gamma=param[2]
        haz=(theta*y)^gamma
        return(haz)
}
c.scale[t.scale<=(3.44)]=1
data.cp.scale.test=cbind(t.scale,c.scale)
cum.haz.cp=function(param,y,a){
        alpha1=param[1]
        alpha2=param[2]
        theta=param[3]
        haz=exp(alpha1*log(theta)+y[,2]*alpha1*log(y[,1])+(1-y[,2])*((alpha1-
                alpha2)*log(a)+alpha2*log(y[,1])))
        return(haz)
```

```
}
cum.haz.gomp=function(param,y){
        beta=param[1]
        gamma=param[2]
        haz=(beta/gamma)*(exp(gamma*y)-1)
        return(haz)
}
cum.haz.gm=function(param,y){
        beta=param[1]
        gamma=param[2]
        alpha=param[3]
        haz=alpha*y+(beta/gamma)*(exp(gamma*y)-1)
        return(haz)
}
x.weib=cum.haz.weib(param.weib,t.scale)
x.cp=cum.haz.cp(param.cp.scale,data.cp.scale.test,a=3.44)
x.gomp=cum.haz.gomp(param.gomp,t.scale)
x.gm=cum.haz.gm(param.gm,t.scale)
cs.weib=rep(0,131)
cs.cp=rep(0,131)
cs.gomp=rep(0,131)
cs.gm=rep(0,131)
cs.weib[status01==1]=x.weib[status01==1]
cs.weib[status01==0]=x.weib[status01==0]+1
cs.cp[status01==1]=x.cp[status01==1]
cs.cp[status01==0]=x.cp[status01==0]+1
cs.gomp[status01==1]=x.gomp[status01==1]
cs.gomp[status01==0]=x.gomp[status01==0]+1
cs.gm[status01==1]=x.gm[status01==1]
cs.gm[status01==0]=x.gm[status01==0]+1
```

Plotting the Cox-Snell residuals for all four models

```
qq.test=rexp(1000,1)
qqplot(cs.weib,qq.test,xlab="Weibull Residuals",ylab="Data Residuals",main="Cox-Snell
        Residual Plot\nWeibull Distribution",pch=16)
abline(0,1,col="red")
qqplot(cs.gomp,qq.test,xlab="Gompertz Residuals",ylab="Data Resdiauls",main="Cox-Snell
        Residual Plot\nGompertz Distribution",pch=16)
abline(0,1,col="red")
qqplot(cs.gm,qq.test,xlab="Gompertz-Makeham Residuals",ylab="Data Resdiauls",main="Cox-
        Snell Residual Plot\nGompertz-Makeham Distribution",pch=16)
```

```
abline(0,1,col="red")
qqplot(cs.cp,qq.test,xlab="Change Point Residuals",ylab="Data Resdiauls",main="Cox-Snell
        Residual Plot\nWeibull Change Point Model",pch=16)
abline(0,1,col="red")
```

References

"Akaike's Information Criterion." *Model Selection*. N.p., n.d. Web. 23 Apr. 2011. <http://www.modelselection.org/aic/>.

"Aluminum Smelting and Refining." *Illinois Sustainable Technology Cente*. University of Illinois, n.d. Web. 12 Mar. 2011. <http://www.istc.illinois.edu/info/library_docs/manuals/primmetals/chapter4.htm>.

Elandt-Johnson, Regina, and Norman Johnson. *Survival Models and Data Analysis*. Wiley-Interscience: John Wiley & Sons, Inc., 1999. 60-63. Print.

"Gompertz–Makeham law of mortality." *Wikipedia, the free encyclopedia*. Wikipedia, n.d. Web. 25 Mar. 2011. <http://en.wikipedia.org/wiki/Gompertz%E2%80%93Makeham_law_of_mortality>.

Hogg, Robert, and Johannes Ledolter. "Ch. 3: Continuous Probability Models."*Engineering Statistics*. New York: Macmillan Publishing Company, 1990. 118-121. Print.

"How Aluminum is Produced."*Rocks, Minerals, Fossils and Earth Science Supplies*. Alcoa Aluminum Institute, n.d. Web. 12 Apr. 2011. <http://www.rocksandminerals.com/aluminum/process.htm>.

Lee, Elisa , and John Wenyu Wang. "Chapter 2: Functions of Survival Time." *Statistical Methods for Survival Data Analysis*. Hoboken: John Wiley & Sons, Inc., 2003. 8-15. Print.

"Likelihood ratio tests."*Information Technology Laboratory Homepage*. NIST/SEMATECH, n.d. Web. 25 Apr. 2011. <http://www.itl.nist.gov/div898/handbook/apr/section2/apr233.htm>.

"Maximum likelihood estimation." *Information Technology Laboratory Homepage*. N.p., n.d. Web. 30 Mar. 2011. <http://www.itl.nist.gov/div898/handbook/apr/section4/apr412.htm>.

O'Meara, Brian. "Model Selection Using the Akaike Information Criterion (AIC) | Brian O'Meara Lab." *Brian O'Meara Lab*. NIST/SEMATECH, n.d. Web. 26 Apr. 2011. <http://www.brianomeara.info/tutorials/aic>.

"Related Distributions."*Information Technology Laboratory Homepage*. NIST/SEMATECH, 1 June 2003. Web. 28 Apr. 2011. <http://itl.nist.gov/div898/handbook/eda/section3/eda362.htm>.

Smith, Peter. *Analysis of Failure and Survival Data*. Boca Raton: Chapman & Hall/CRC, 2002. Print.

Steenbergen, Marco R. “Maximum Likelihood Programming in R.” Chapel Hill: 2006.

"Weibull Distribution."*Information Technology Laboratory Homepage*. NIST/SEMATECH, n.d. Web. 2 Feb. 2011. <http://itl.nist.gov/div898/handbook/eda/section3/eda3668.htm>.